

**TASK**

**Exploratory Data Analysis on the \*\*\*\*\*\*\*\*\*\*\*\* Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Summary of the data set

**DATA CLEANING**

# SUMMARY OF THE METHODS AND VISUALIZATIONS DONE DURING DATA CLEANING

MISSING DATA

# ANY MISSING DATA? HOW DID YOU HANDLE IT

DATA STORIES AND VISUALIZATIONS

# THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALIZATIONS OF THE DATA

# ENSURE THIS DOCUMENT IS NEAT AND CAN BE ADDED IN YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY : ABRAHAM WORDOF**



**The Medicine dataset**

The Medicine dataset consisted of 19 columns, namely:

['applicants\_mcc\_licence\_no', 'applicant\_name', 'mcc\_medicine\_reg\_no', 'nappi\_code', 'atc\_4', 'medicine\_schedule', 'medicine\_proprietary\_name', 'active\_ingredients', 'strength', 'unit', 'dosage\_form', 'pack\_size', 'quantity', 'manufacturer\_price', 'logistics\_fee', 'vat', 'sep', 'unit\_price', 'originator\_or\_generic']

**DATA CLEANING AND MISSING DATA**

After just running the first 5 rows of the data set, we see that they are many null values. Next let us plot a heatmap to check for all null values. So, I then plotted a heatmap to get an idea of how many null values the data set consists of. The heatmap shows us that the dataset contains many null values and because it is a big dataset, and the data is MCAR we can drop all the rows with the null values. We can start by dropping the "effective\_date", "status" and "\_12\_months\_volume\_of\_sales" columns because they contain the most with null values, then go on further and drop all rows with null values.

DATA STORIES AND VISUALIZATIONS

* I started by creating a heatmap of all the columns with numerical values. The above heatmap shows us how the data is distributed in all the columns and which columns have the highest values. The Nappi code, SEP and Manufacturer price have the highest values in the dataset based solely on looking at the above visualization.
* Next, I created a box plot visualization of the Numbers of Companies and how many times they applied for medication. The visualization shows us how many companies they are and also how many times they have applied to receive medication. The most common company fount in this visualization is Adcock Ingram Care which has more than 170 applications followed by Accord Healthcare with more than 110 applications and finally Abbott Laboratories with more than 70 applications.
* I then created 2 box plot visualizations, one for the manufacturer price and one for the unit\_price of the different medications. The first box plot shows us that the most common manufacturer price is mostly between 80 and 260 with multiple outliers ranging from 550 and even a maximum of 1000. The second box plot shows us that on average the unit price is very cheap and below 30. The box plot shows multiple outliers ranging from 50 to a maximum of about 820.
* From there, I then created a 2 regplot which is basically a scatterplot and a linear regression in one. The first regplot plots the relationship between the Unit price and manufacturer price. The visualization shows that they is a direct relationship between the Unit Price and Manufacturer Price. The higher the Unit Price the higher the Manufacturer Price. The second regplot shows the relationship between Unit Price and VAT and just like the Unit Price and Manufacturer price the Unit Price and the Vat price have a direct relationship.
* The next visualization created was a box plot for the SEP of medicines. The average SEP for the dataset ranges from 100 to about 300. The SEP column also contains multiple outliers with the highest values ranging from 800 to about 1000.
* A count plot for the number of Originator and Generic medications was then created. This visualization contains only 2 different types of medication (Originator and Generic medications). The two types of medication have a similar count with the originator having a slightly higher value than the generic mediation.
* From there I plotted a bar graph for the top 10 most active ingredients. Because we have a very big dataset, I have only plotted the top 10 highest values in the Active Ingredients column. The Calcium Chloride is the most common ingredient and is found in more than 25 medications followed by glucose which is found in about 17 different types of medication.
* Next, I plotted a bar graph of the top 10 most common medicines. The 4 most common medications are Hytrim, Sabax Dianeal 2.5, Sabax Dianeal LS 4.25 and Sabax Dianeal Normal Calcium with 1.5%. The next 5 most common medications are Nexiflux 40, Dehrin Solution, Mavik (was Gopten) and Nexiflux 20.
* A bar graph of the top 5 most common dosages of medicine. TAB is the most common medication dose found more than 120 times followed by INJ found more than 90 times and INF found more than 80 times.
* A pie chart of medicine schedules. The pie chart shows the most common medicine schedule. The S3 and S4 are the most common medicine schedules, and both have a combined value of over 75% of the whole pie chart, with each having a value of about 37.5%. Followed by S1 with about 10% of the pie chart.
* A box plot for the logistic fees. This box plot visualization shows us that the average logistic fees are less than 50. Just like the unit price box plot, this visualization shows us that there are also multiple outliers ranging from 100 to 1000.